**DECLARATION OF ERIC BRENNER IN OPPOSITION TO PLAINTIFFS' MOTION IN LIMINE TO EXCLUDE THE LITIGATION SURVEY BY DEFENDANTS' EXPERT KEVIN LANE KELLER, PH.D.**

# EXHIBIT 9

Annu. Rev. Sociol. 2003. 29:65–88
doi: 10.1146/annurev.soc.29.110702.110112


# THE SCIENCE OF ASKING QUESTIONS

Nora Cate Schaeffer[1] and Stanley Presser[2]
[1]*Sociology Department, University of Wisconsin, Madison, Wisconsin 53706; email: Schaeffe@ssc.wisc.edu*
[2]*Sociology Department, University of Maryland, College Park, Maryland 20742; email: Spresser@socy.umd.edu*

**Key Words** questionnaires, survey research, measurement, methodology, interviewing

■ **Abstract** Survey methodologists have drawn on and contributed to research by cognitive psychologists, conversation analysts, and others to lay a foundation for the science of asking questions. Our discussion of this work is structured around the decisions that must be made for two common types of inquiries: questions about events or behaviors and questions that ask for evaluations or attitudes. The issues we review for behaviors include definitions, reference periods, response dimensions, and response categories. The issues we review for attitudes include bipolar versus unipolar scales, number of categories, category labels, don't know filters, and acquiescence. We also review procedures for question testing and evaluation.

## INTRODUCTION

Research on the wording of survey questions flourished in the first two decades after the modern sample survey was invented, culminating in Stanley Payne's 1951 classic, *The Art of Asking Questions*. With the notable exception of research on acquiescence, attention to wording then waned over the next quarter of a century. In the past two decades, there has been a revival of interest by survey methodologists, who have drawn on and contributed to work by cognitive psychologists, conversation analysts, and others to lay a foundation for the science of asking survey questions.

The standardized survey interview is a distinct genre of interaction with unique rules, but it shares many features with ordinary interaction because social and conversational norms as well as processes of comprehension, memory, and the like are imported into the interview from the situations in which they were learned and practiced. As a result, contributions to the science of asking survey questions also enhance our understanding of other types of interviews and of social interaction in general—many processes can be studied "in surveys as in life" (Schuman & Ludwig 1983).

Methodologists have applied information-processing models from cognitive psychology to explain how questions are answered in survey interviews

(Sirken et al. 1999, Sudman et al. 1996, Tourangeau et al. 2000), and these models have influenced much of the research that we review here. At the same time, there has been renewed attention to how the interviewer and respondent interact (Schaeffer & Maynard 1996, Maynard et al. 2002). There is an intricate relationship among the survey question as it appears in the questionnaire, the rules the interviewer is trained to follow, the cognitive processing of the participants, the interaction between the interviewer and respondent, and the quality of the resulting data. In an interviewer-administered survey, the question that appears on the screen or the page may be modified in the interaction that ultimately produces an answer, and in a self-administered survey, conventions learned in other social contexts may influence how a respondent interprets the questions presented (e.g., Schwarz 1994). Nevertheless, we proceed here as though the text of the question the respondent answers is the one that appears in the questionnaire, although in one section we review recent experiments that modify the traditional practices associated with standardization.

Researchers must make a series of decisions when writing a survey question, and those decisions depend on what the question is about. Our review is structured around the decisions that must be made for two common types of survey questions: questions about events or behaviors and questions that ask for evaluations or attitudes. Although there are several other types of questions (e.g., about knowledge and sociodemographic characteristics), many survey questions are of one of these two types.

In some cases, research suggests an approach that should increase the reliability or validity of the resulting data, for example, labeling all the categories in a rating scale. In other cases, the literature only suggests how different design alternatives, such as using a checklist instead of an open question, lead to different results without clearly showing which approach is best or even clearly specifying what best means.

Researchers who compare different ways of asking standardized questions use various methods to evaluate the results. The traditional approach involves split-sample experiments, which sometimes include measures of reliability (split-half or over time) and validity (construct, convergent, or discriminant). Other approaches that have increasingly been used include cognitive evaluation or expert review, feedback obtained from respondents during cognitive interviews or debriefing questions, the results of coding the interaction between interviewers and respondents, and feedback from interviewers in debriefing sessions (see Testing and Evaluating Questions, below).

The interactional mode (interviewer or self-administered) and technological mode (computer or paper) influence the nature of both a survey's questions and the processes used to answer them. For example, a household roster may look different and be administered differently depending on whether it is implemented using a grid on paper or a more linear implementation on a computer (Moore & Moyer 1998, Fuchs 2002). Seeing the questions in a self-administered form rather than hearing them read by an interviewer, to take another example, may

mitigate the effects of question order or make it easier for respondents to use the full range of categories in rating scales (Bishop et al. 1988, Ayidiya & McClendon 1990, Dillman & Mason 1984). Nevertheless, investigators face similar decisions regardless of the mode, and many of the topics we discuss have been examined in several modes.

## THEORETICAL APPROACHES

Models of the process of answering a survey question vary in the number of stages they present, but most models include understanding a question, retrieving or constructing an answer, and reporting the answer using the specified format (Cannell et al. 1981, Tourangeau 1984, Turner & Martin 1984, Sudman et al. 1996). Research that draws on these models has generally focused on the first two stages. Respondents may use somewhat different methods of processing for questions about events or behaviors than they do for questions that request evaluations or attitudes.

As respondents hear or read a survey question, they construct a "pragmatic meaning" that incorporates their interpretation of the gist of the question, why it is being asked, and what constitutes an acceptable answer. Using an early version of what we would now call cognitive interviews, Belson (1981) described how respondents may expand or restrict the meaning of the concepts in a question. Research based on information-processing models has identified some of the mechanisms by which this occurs. For example, conversational norms specify that each contribution to a conversation should convey new information. Thus, if respondents are asked to rate first their marriages and then their lives as a whole, they may interpret the second question as being about their lives other than their marriages (Schwarz et al. 1991, Tourangeau et al. 1991). Because they have already conveyed information about their marriages, they provide only new information when rating their lives as a whole. Respondents may also expand or restrict the meaning of concepts because the wording of a question evokes prototypes or exemplars that then dominate the definition of the concept. Thus, even though a researcher words a question to ask about health practitioners, a respondent may still think the question is about physicians because that prototypical health practitioner is so salient. The respondent's interpretation of the question can also be influenced by information in response categories. In a 1986 study, for example, many respondents selected the invention of the computer as an important social change when the category was offered in a closed question, but it was seldom mentioned in response to an open question (Schuman & Scott 1987).

Most cognitive psychologists believe that memories and autobiographical reports of events are as much constructed as retrieved. Thus, if respondents can easily think of an instance of an event, they may infer that the event was frequent—an instance of the availability heuristic (Tversky & Kahneman 1973). Similarly, if a memory is vivid, the event may be reported as occurring more recently than it did (Brown et al. 1985, Bradburn et al. 1987). Respondents who believe they have

changed (or stayed the same) may be guided by that belief when recalling the past (Ross & Conway 1986). Efforts to remember may also focus on salient prototypes or exemplars, such as physicians instead of health practitioners, and less-salient incidents may be omitted.

In producing answers about the frequency of events, respondents use a variety of methods, including counting episodes, rate-based estimation (either for a class of events as a whole or decomposing and estimating for members of the class), relying on cues conveyed by the response categories, guessing, and various combinations of these (Blair & Burton 1987). Estimation strategies lead to heaping at common numbers, such as multiples of 5 or 10 (Huttenlocher et al. 1990). Many of these strategies can be considered techniques for "satisficing," that is, for conserving time and energy and yet producing an answer that seems good enough for the purposes at hand (Krosnick 1991). These examples also illustrate the important point that comprehension and the retrieval and construction of an answer are not completely sequential or independent processes.

As powerful as information-processing models have been in helping us understand how survey questions are answered, they can usefully be supplemented by paying attention to the social aspects of surveys. Some of what we might otherwise label cognitive processing (if we did not look at the behavior of the participants) actually occurs in the interaction between the interviewer and respondent (Schaeffer & Maynard 1996, Maynard et al. 2002). For example, when respondents hesitate or provide answers that do not match the format of the question, interviewers may use this information to diagnose that the respondent does not understand something about the question and thus intervene (Schaeffer & Maynard 2002). Moreover, the reporting task a respondent confronts may be affected by the respondent's value on the characteristic being reported, which is usually bound up with (if not determined by) social factors. For example, a respondent with a complicated employment history will find it difficult to report beginning and ending dates of jobs, whereas this task will be simpler for someone who has held the same job since completing school. Information about the distribution of employment experiences will assist researchers in anticipating the different response strategies respondents are apt to adopt and therefore the different errors they will make in answering questions about employment history (Mathiowetz & Duncan 1988, Schaeffer 1994, Dykema & Schaeffer 2000, Menon 1993). The fact that true values and the errors made in measuring those values are functions of the same social processes also means that the assumptions of many statistical models may often be violated (Presser & Traugott 1992).

## QUESTIONS ABOUT EVENTS AND BEHAVIORS

Surveys ask about an astonishing variety of events and behaviors, from using automatic teller machines, caring for children, and visiting physicians to using contraception and voting in elections. In addition, many questions that do not initially appear to be about events actually do concern characteristics of events.

For example, a question about the total amount of wages received in the last month implicitly refers to the events of working and being paid, and "household composition" is a function of who stays, eats, or receives mail in a place.

The first consideration in asking about an event or behavior is whether members of the target population are likely to have encoded the information. For example, researchers may want to ask parents about their children's immunization history, but some parents will be unaware of what, if any, immunizations their children have received (Lee et al. 1999). For events that respondents do encode, two major types of errors affect self-reports. Omissions result when individual events are forgotten because of dating errors (e.g., when events are "telescoped" backward in time and so are incorrectly excluded from the reference period), because similar events become conflated in a "generic memory" (Means et al. 1989), or because the wording of a question leads the respondent to search some areas of memory (e.g., visits to physicians) while neglecting others (e.g., visits to nurse practitioners). By contrast, intrusions result when events are telescoped forward in time (and are incorrectly included in the reference period) or when memories are altered by scripts, schemata, or embellishments from retellings over time.

Researchers must determine the level of accuracy they will try to achieve with the analytic goals and resources at hand. Many techniques that hold promise for improving the accuracy of self-reports require additional interviewing time or additional resources for questionnaire development, testing, and interviewer training. Interviewers must also be able to implement the methods and respondents willing to tolerate them.

## Naming and Defining the Event

The close relationship between comprehension and retrieval is evident in the most basic technique for improving retrieval—providing a cue in the question that matches the label under which the respondent has stored information about the event. Focus groups and other developmental interviews can identify the vocabulary respondents use, that is, they can help investigators map analytic constructs onto native constructs (Schaeffer & Dykema 2003). Not surprisingly, attempts to use the language of the respondent to provide cues immediately encounter problems. The most obvious is that different groups within the target population may use different vocabularies or organize their behavior in different ways.

There are two general approaches to defining a target event or behavior. One is to provide the analytic definition and ask the respondent to answer in its terms. The other is to allow respondents to answer using their native concepts and structure the sequence of questions so that the analyst can map native concepts onto the analytic concept. The contrast can be seen by comparing several versions of a question about vehicles. The first version provides a definition by listing the members of the category:

> Next I'd like to know about vehicles, including automobiles, motorcycles, motor scooters, and trucks. How many vehicles do you own?

The next version uses a checklist to implement the definition and asks about automobiles last to prevent misclassification due to respondents thinking of vehicles in other categories as automobiles:

> I'm going to read you a list of different types of vehicles. As I read each one, please tell me how many vehicles of that type you own. How many trucks do you own? Motor scooters? Motor cycles? Automobiles?

The third example allows respondents to answer using their own concept and then asks about vehicles they might have omitted:

> How many vehicles do you own? IF ONE: Is that an automobile, truck, motor scooter, or motor cycle? IF MORE THAN ONE: How many of them are trucks? Motor scooters? Motor cycles? Automobiles? In addition to the vehicle(s) you just told me about, do you own any (LIST TYPES OF VEHICLES NOT MENTIONED)?

Providing a definition is probably most appropriate for events that can be defined simply (or for questions that ask respondents to classify themselves with respect to some social category with which they are familiar and that has a well-known name). When a definition is provided, it should precede the actual question. If the definition follows the question, interviewers will frequently be interrupted before the definition is read, which will lead both to an increase in interviewer variance (as interviewers handle the interruption differently) and to not all respondents hearing the definition (Cannell et al. 1989, Collins 1980).

Investigators sometimes include examples as part of the definition to clarify the concept. Respondents will focus on those examples when they search their memories. For a complex and heterogeneous class of events (for example, arts events the respondent has participated in or attended), a checklist that asks separately about each member of the class is often used. Checklists appear to reduce omissions, probably because recognition tasks are easier than free recall tasks and because the list structure requires that respondents take more time to process each item. On the other hand, checklists may increase reporting of events that took place before the reference period, and they may lead to overestimates for small categories if the event class is "decomposed" inappropriately (Belli et al. 2000, Menon 1997). Thus, a checklist is apt to yield higher overall levels of reporting for a class of events than a single question about the class that includes examples.

The definition of a complex event can often be unpackaged into a series of simpler items, each of which asks about a component of the definition. Consider the following item:

> During the past 12 months since July 1st 1987, how many times have you seen or talked with a doctor or a medical assistant about your health? Do not count any times you might have seen a doctor while you were a patient in a hospital, but count all the other times you actually saw or talked to a medical doctor of any kind about your health.

It can be revised as follows:

> Have you been a patient in the hospital overnight in the past 12 months since July 1st 1987?
>
> (Not counting when you were in a hospital overnight) During the past 12 months since July 1st, 1987, how many times did you actually see any medical doctor about your own health?
>
> During the past 12 months since July 1st 1987, were there any times when you didn't actually see the doctor but saw a nurse or other medical assistant working for the doctor?
>
> During the past 12 months since July 1st 1987, did you get any medical advice, prescriptions, or results of tests over the telephone from a medical doctor, nurse, or medical assistant working for a doctor? (Cannell et al. 1989, appendix A, p. 1).

## Reference Periods

The choice of reference period is usually determined by the periodicity of the target events, how memorable or patterned the events are likely to be, and the analytic goals of the survey. Thus, investigators may ask about religious practices over the previous year to obtain information about respondents who attend services only on their religion's (annual) holy days. By contrast, questions about purchases of candy bars usually use a much shorter reference period. Although more recent events are generally remembered better than more distant events, the influence of the length of the reference period is probably smaller for frequent and highly patterned events, presumably because respondents use information about patterning to construct their answers (Schaeffer 1994, Dykema & Schaeffer 2000).

Researchers must decide how (and how often during a series of questions) to specify the reference period they have selected. Schaeffer & Guzman (1999) found only weak evidence in support of their prediction that using more specific boundaries (e.g., specifying the start and end of the reference period) would reduce telescoping and lead to lower levels of reporting. The reference period may also influence how a question is interpreted. For example, a question about how often the respondent has been angry in the past year is interpreted as referring to more intense episodes than a question that refers to the past week, but the informational value of the reference period is attenuated when it is repeated for many questions (Winkielman et al. 1998, Igou et al. 2002). Experiments with "anchoring" techniques, in which respondents are shown a calendar that has the reference period marked on it and asked to think of events that occurred within that time frame, have sometimes resulted in higher levels of reports of threatening behaviors as well as improved internal consistency of the reports (Turner et al. 1992, Czaja et al. 1994).

We suspect that the reference period should usually be given at the beginning of a question (so that respondents do not construct their own before hearing the investigator's) and that it should be fully specified at the beginning of a line of

questioning and then given in abbreviated form, and in a parallel location, in subsequent questions. For example, a question that introduces a topic might say the following:

> The next questions are about the amount of child support you actually received between January 1 and December 31, 2001. In 2001, did you receive any payments for child support?

Subsequent questions would then use the abbreviated specification of the reference period:

> In 2001, how many payments for child support did you receive?

The repetition communicates that the reference period has stayed the same; using the abbreviated form and the parallel structure conserves cognitive processing.

Overreporting errors due to forward telescoping can be reduced using a panel survey with bounded interviews (Neter & Waksberg 1964). In an initial bounding interview, respondents are asked to report events; the list of events reported at time 1 is then consulted during the time 2 interview. If an item is reported at time 2, the interviewer verifies that it is a new item and not a duplicate of the item reported at time 1.

It may also be possible to obtain the effects of "bounded recall" with a single interview by asking about two reference periods, one of which includes the other. Reductions in reporting that are consistent with those observed for true bounded recall have been observed by asking first about the past 6 months and then about the past 2 months (Loftus et al. 1992, Sudman et al. 1984).

When collecting information about related events over long periods of time, event history calendars are useful (Freedman et al. 1988, Means & Loftus 1991). In a new implementation of the methodology that draws on recent advances in theories of memory, the calendar appeared to improve reporting about several variables—such as moves, income, and weeks unemployed—although it increased overreporting of other variables (Belli et al. 2001).

## Modifying Standardization to Improve Accuracy

Respondents who are uncertain about the intent of a question may ask the interviewer for clarification. Yet the dictates of standardization (motivated by a concern that all respondents hear the same information) mean interviewers are usually not allowed to provide substantive help. Critics of standardization have pointed to this as a weakness of the traditional approach to survey interviewing (e.g., Suchman & Jordan 1990), and research has begun to explore more flexible approaches to interviewing. Oksenberg et al. (1992) experimented with a questionnaire structure and interviewing procedure designed to improve the respondent's recall, understanding of survey concepts, and motivation to work hard. Their interviews began with a free-flowing discussion of the topic, made extensive use of a calendar and timeline, and authorized interviewers to design probes using information they learned

during the interview. As far as we know, however, these methods have yet to be evaluated or implemented on a large scale.

Schober & Conrad (1997) (Conrad & Schober 2000) have experimented with a less-radical approach. In experiments with short questionnaires administered by telephone, they found that allowing interviewers to depart from the question-wording in an attempt to ensure that respondents correctly understood the questions improved the reporting of consumer purchases (although at the cost of lengthier interviews). The feasibility of this approach for other kinds of surveys is uncertain. For surveys covering a wide selection of topics, the range of concepts the interviewer must understand will make it more challenging to provide accurate clarifications. In longer interviews, it may be harder to sustain respondent motivation to request clarification when it is needed. For in-person surveys, decentralized administration may make it impractical to monitor interviewer clarifications, and for large-scale surveys, the size of the interviewing staff will make it more difficult to ensure uniform interviewer understanding of question intent.

## Response Dimensions and Response Categories

Several response dimensions are relevant for questions about events: occurrence, absolute frequency (the number of times the event occurred within the reference period), relative frequency (how often it occurred, using adverbial phrases such as "very often"), regularity or periodicity, and date of occurrence. Investigators are often interested in only one or two of these dimensions, and respondents probably will not tolerate being asked about all of them. Question forms and response categories can be built around these response dimensions in various ways.

Estimates of occurrence, the proportion who engaged in the behavior within the reference period, can be obtained in three different ways: simple yes-no items, closed frequency items, and open frequency items. If its lowest category makes clear that any occurrence should be reported, then a closed frequency question may yield higher estimates of behavior than a yes-no item, possibly because respondents with low frequencies see the yes-no inquiry as asking whether they are more similar to those who engage in the behavior or to those who do not (J. Dykema & N.C. Schaeffer, unpublished work). Open frequency inquiries may make it even more likely that a respondent will report experiencing an event because the question may be interpreted as presuming the event occurred (Knauper 1998), although J. Dykema & N.C. Schaeffer (unpublished work) did not find this effect.

When it comes to absolute frequencies, open questions may obtain higher estimates for threatening behaviors among those who have engaged in the behaviors (Blair et al. 1977). Because such behaviors are underreported, this suggests they are more accurately measured by open questions than closed questions. Closed questions also have the disadvantage that respondents may be influenced by where the boundaries are drawn to create the response categories. For example, Schwarz et al. (1985) found that estimates of television viewing depended on whether the lowest response category suggested that respondents watch a lot (up to 2.5 hours) or a little (up to .5 hours). Tourangeau & Smith (1996) reported a similar effect for

measures of number of sex partners. Likewise, when the response categories suggested that "being really annoyed" was rare, respondents generated more extreme examples of being really annoyed than they did when the response categories suggested that the event might be common, which implies that their interpretation of the question had been affected by the response categories (Schwarz et al. 1988). Although Burton & Blair (1991) did not find any difference in the accuracy of open and closed questions about writing checks and using automatic teller machines, and open questions do not always obtain higher frequencies for threatening behaviors (e.g., Tourangeau & Smith 1996), the potential hazards of closed questions means that open questions are usually preferable (Schaeffer & Charng 1991).

Questions about relative frequencies use vague quantifiers, such as "very often, pretty often, not too often, seldom, never" (Bradburn & Miles 1979; for a recommended set, see Pohl 1981). Relative frequencies are not simple translations of absolute frequencies; they incorporate evaluative information. As a result, conclusions about group differences may vary depending on whether one examines absolute or relative frequencies (Schaeffer 1991). This is nicely illustrated in Woody Allen's *Annie Hall*. Both Annie and Alvie Singer report that they have sex three times a week, but she characterizes this as "constantly," whereas his description is "hardly ever." In addition to conveying information about preferences or expectations, relative frequencies may express how respondents compare themselves with similar others. Relative frequencies are probably most appropriate when the investigator wants to give weight to the evaluative component in the respondent's perception of the frequency, when group comparisons are not a central analytic goal, or when the frequencies are too difficult to report in an absolute metric.

Even absolute frequencies (which, like all self-reports, contain errors) may include evaluative information similar to that in relative frequencies. Which frequency, absolute or relative, is cognitively prior probably differs for different events and for different patterns of events. In some cases, a respondent who is offered response categories that express relative frequency may retrieve an absolute frequency from memory that must then be translated into the relative metric, whereas in other cases, a respondent who is offered an absolute reporting format may retrieve a relative frequency and then translate it to the absolute metric (Conrad et al. 1998).

## Issues of Relevance

Questions are usually written in a way that presupposes they are relevant to the respondent. Consider the following question: In the week beginning last Sunday and ending yesterday, how many hours did you work outside in your garden? Because not all respondents have a garden, a category labeled "IF VOLUNTEERED: Respondent does not have a garden" must be provided for interviewers. With the exception of behaviors known to be underreported (and those engaged in by almost everyone), it is better to avoid this kind of question. Respondents may be annoyed at being asked a question that does not apply to them, and interviewer variability may be higher for items with if volunteered categories (Collins 1980). Even

in self-administered questionnaires where the category is visible, it may present problems. The if volunteered category can be described as a hidden question and usually should be replaced by a filter question that establishes the relevance of a subsequent line of questioning (Forsyth et al. 1992).

Conditional events, those that are sometimes relevant, pose special challenges. In an example provided by Fowler (1992), one third of respondents had difficulty answering the following question: What is the number of servings of eggs you eat on a typical day? Respondents had fewer problems when the question was revised as follows: On days when you eat eggs, how many eggs do you usually have? The latter version recognizes the difficulty of averaging across days when the respondent did and did not eat eggs and is thus easier to answer.

### Threatening Behaviors

In the past decade, there was considerable experimentation with methods for improving the accuracy of reports of socially undesirable behaviors. These studies focused on drug use, abortions, sexual behavior, and (non)voting, and some of them used checks of external records to evaluate the experimental methods (see reviews in Schaeffer 2000 and Tourangeau et al. 2000). The results have been mixed. For example, wording changes that tried to make respondents more comfortable admitting that they did not vote did not reduce reports of voting (Presser 1990, Abelson et al. 1992), and similar attempts to increase the reporting of abortions have also failed (e.g., Jobe et al. 1997). Belli et al. (1999), however, were able to reduce voting claims with a question that asked the respondent to remember details about the vote, presented multiple response categories for "didn't vote" (e.g., "I thought about voting this time but didn't"), instead of the usual single one, and phrased the "did vote" category in definite terms ("I am sure I voted in the November 5 election"). Further work is needed to identify which of these changes is key. The most consistent finding in this literature is that more private (e.g., self-administered) modes of administration produce both higher reports of socially undesirable behaviors (Tourangeau & Smith 1996, Turner et al. 1998) and lower reports of socially desirable ones (Presser & Stinson 1998).

## QUESTIONS ABOUT SUBJECTIVE PHENOMENA

For questions about events and behaviors, error can be thought of as the difference between the report of a respondent and that of an omniscient observer. This conception of a "Platonic true score" (Bohrnstedt 1983) does not apply to measures of subjective phenomena, but it is still useful to think of error in attitude questions in terms of sources of response variation other than the target construct. As with questions about events and behaviors, these sources include comprehension problems, lack of motivation to answer carefully, and response sets such as acquiescence. Many design decisions for subjective items—for example, how to label the midpoint of a bipolar scale—have implications for how a construct is conceptualized.

Similarly, how an investigator conceptualizes a construct may suggest some design choices, such as whether or not to use a filter question.

## Basic Structure: Objects and Evaluative Dimensions

Questions about subjective phenomena have two basic components: an object and an evaluative dimension. For example, a respondent might be asked to express approval or disapproval (evaluative dimension) of the Agricultural Trade Act (object), or to rate himself (object) on happiness (evaluative dimension). The content of the questions and the constructs they measure vary significantly and include questions about norms (e.g., Do you agree or disagree that adult children should care for their aging parents?), support for policies (e.g., Do you favor or oppose President X's policy about Y?), and internal experiences or states (e.g., How sure are you that you want to have another child?).

The first decision an investigator faces in writing a subjective question is selecting names or labels for the object and the evaluative dimension. The goal is to select names that are easy to understand and that will be understood similarly by all respondents. In addition, the evaluative dimension must seem appropriate for the object. For example, respondents would likely think questions about how "beautiful" they find televison news programs weird and off-putting. Focus groups are often used during questionnaire development to identify appropriate names and relevant dimensions.

## Basic Question Forms

Subjective questions take various forms, including ratings, rankings, agree-disagree statements, forced choices between statement pairs, and open-ended inquiries. Some research suggests that rankings—for example, ordering the importance of a set of job characteristics—have advantages over ratings—for example, evaluating the importance of each job characteristic using a common scale (Krosnick & Fabrigar 2003). However, rankings can only be used with a small number of objects, particularly in telephone interviews, and they also take more interview time and pose special analysis problems. As a result they are not commonly used. For similar reasons, open questions asking for reasons or objects (e.g., important traits for children to possess) are also relatively uncommon. Although open questions can be indispensable in certain circumstances (e.g., to document the absence of a response), they are much more expensive than closed questions (requiring substantially more interview time and postinterview processing) and are more difficult for respondents to answer and for researchers to analyze.

## Decisions for Ratings: Bipolar Versus Unipolar

Rating scales can be structured as either bipolar (e.g., extremely boring to extremely interesting) or unipolar (e.g., not at all interesting to extremely interesting). The bipolar scale has a midpoint at which there is a transition (e.g., from boring to

interesting). This midpoint can be conceived of as indicating either indifference (e.g., neither boring nor interesting) or ambivalence (e.g., boring in some ways and interesting in others), so that the definition of the midpoint potentially affects the meaning of other points as well. One might assume that the category "not at all interesting" in the unipolar version includes all the positions between "extremely boring" and "neither boring nor interesting" in the bipolar version, but little is known about how respondents actually perceive the difference between the two versions. A potential disadvantage of bipolar items is that they assume more about the evaluative continuum than unipolar items do, for example, that the poles are, in fact, opposites. Indeed in some cases, it may be challenging to find appropriate opposites to use to label the endpoints of a bipolar scale. Unipolar items make fewer assumptions, but they risk irritating respondents who see questions that present negative and positive dimensions separately as repetitive or inappropriate.

Bipolar (and unipolar) scales are often presented, either orally or on a showcard, with verbal labels for endpoints and numeric labels for the intervening categories. One might expect that when bipolar verbal labels are combined with bipolar numeric labels (e.g., $-5$ to $+5$ versus 0 to 10), they would reinforce each other and appear clearer to respondents than other combinations of verbal and numeric labels. Nonetheless, verbal and numeric labels appear to have separate effects that do not interact. Bipolar numeric labels move the response distribution toward the positive end when compared with unipolar numeric labels, and bipolar verbal labels result in more use of the middle category and less use of the negative pole when compared with unipolar verbal labels (O'Muircheartaigh et al. 1995). In experiments with bipolar items reported by Schaeffer & Barker (1995), ad hoc scales made by combining items had higher reliability when the items had unipolar numeric labels ranging from 1 to 7 rather than bipolar labels ranging from $-3$ to $+3$ when the topic was approval of government, but the numeric labels had little impact for questions about economic satisfaction. Unipolar numeric labels that begin with 0 are also likely to be interpreted differently from those that begin with 1 (Schwarz et al. 1998).

## Decisions for Bipolar Scales

Several design decisions are unique to bipolar rating scales: whether to use an "unfolding" or "branching" technique, whether to offer a middle category, and how to label the middle category if it is offered. In unfolding (see Groves & Kahn 1979), the respondent is first asked about valence or direction (e.g., Overall, are you satisfied or dissatisfied?) and then about degree (e.g., Are you only a little satisfied, slightly satisfied, somewhat satisfied, very satisfied, or extremely satisfied?). Krosnick & Berent (1993) showed that a composite based on a pair of fully labeled branching (unfolded) items took less time in the interview, produced more consistency in a reinterview, and predicted criterion variables better than a single, partially labeled nonbranching question. The extent to which the results were due to labeling versus branching is unclear. Using a three-wave panel, Alwin (1992) estimated that the reliability of a composite measure of party identification was

only slightly greater than the reliability of the valence component and modestly greater than the intensity component. The reliability of the composite was substantially greater than the reliability of the other 7-point scales in that analysis, but, as in Krosnick & Berent, the items compared differed in extent of labeling (as well as in content). Although the evidence that fully labeled unfolded items increase reliability is not clear-cut, such items have the advantage of providing a large number of categories without a showcard, which means they can be implemented in the same way in face-to-face and telephone surveys, making them very useful for mixed mode designs and comparisons across modes.

Researchers have long known that when a middle category is offered it will be chosen by more respondents than will volunteer that answer when it is not offered (Schuman & Presser 1981). O'Muircheartaigh et al. (1999) concluded that offering a middle alternative in rating scales reduces the amount of random measurement error and does not affect validity. For some constructs, the label used for the middle category may affect how often it is chosen, e.g, when they rated capital punishment, more subjects chose the middle category when it was labeled "ambivalent" than when it was labeled "neutral" (Klopfer & Madden 1980). This appears to be true whether the scale uses unipolar or bipolar numeric labels (Schaeffer & Barker 1995).

## Category Labels and Intervals Between Categories

When there are only numeric labels for the categories between the verbally labeled endpoints, respondents probably assume the categories are equidistant (Klockars & Yamagishi 1988). However, providing verbal labels for all the categories, both endpoints and intermediate ones, produces more reliable measurement (Alwin & Krosnick 1991). To select verbal labels that define relatively equidistant categories, investigators can refer to studies that scale adverbial expressions of intensity, amount, and likelihood (Cliff 1959, Dobson & Mothersill 1979, and additional sources referenced in Schaeffer 1991). For example, averaging across Cliff's three samples, one might select "not at all, slightly, somewhat, pretty, very, and extremely" as a set of labels. These studies suggest that "very," which commonly appears as an anchor, is probably not intense enough for most applications. However, the studies have generally relied on small, nonrandom samples; thus, replication of their results with larger probability samples of the general public would be desirable.

## Number of Categories

The choice of the number of categories represents a compromise between the increasing discrimination potentially available with more categories and the limited capacity of respondents to make finer distinctions reliably and in similar ways. Based largely on psychophysical studies, the standard advice has been to use five to nine categories (Miller 1956, Cox 1980), although even that number of categories can be difficult to administer in telephone interviews. Both Alwin

& Krosnick (1991) and Alwin (1992) found evidence that the reliability of individual rating scales appeared to increase as the number of categories grew, up to approximately seven or nine categories, with the exception that reliability was greater with two than three categories. Their results must be interpreted cautiously, however, because the questions that were compared differed not only in the number of categories, but also in a large variety of other ways. In a comparison that controlled item content, 11-point feeling thermometers showed higher reliabilities than 7-point scales (Alwin 1997), but the results may have been due to order of presentation, as respondents always answered the feeling thermometers after the rating scales.

A few response methods, such as magnitude scaling and feeling thermometers, offer a very large number of numerical options, but respondents usually choose answers that are multiples of 5, 10, or 25 (at the low end of the continuum) and 50, 100, or 1000 (at the higher end), so that the number of categories used is less than one might expect (Tourangeau et al. 2000). Because most of the categories are unlabeled, respondents' interpretations of them may vary, although assigning labels to a subset of the categories (as is often done with feeling thermometers) probably causes further clustering of answers (Groves & Kahn 1979; Alwin & Krosnick 1991, p. 175, footnote 11). In addition, some respondents find these response tasks difficult, so the proportion who refuse to answer or say they do not know is substantially higher than with other rating scales (Schaeffer & Bradburn 1989, Dominitz & Manski 1997).

## Issues of Relevance: The "Don't Know" Category

The typical attitude item (e.g., Do you favor or oppose X?) implicitly assumes the respondent holds an opinion and therefore may communicate the expectation that a position should be chosen. Experiments have shown that "quasi filters," which explicitly mention "no opinion" as a response option, can substantially lower the number of respondents offering opinions. Moreover, "full filters," which initially ask an entirely separate item about whether one has an opinion, reduce the proportion offering opinions even more (Schuman & Presser 1981).

These findings have led some researchers to recommend filters as a solution to the problem Converse (1964) described as "nonattitudes," answers from respondents with no opinion that are arrived at through a process akin to mentally flipping coins. However, the responses Converse interpreted as nonattitudes were elicited by questions preceded by full filters that clearly legitimated the expression of no opinion. (In addition, respondents were told that not everyone was expected to have an opinion.) Thus, filters are unlikely to solve the problem Converse diagnosed.

It is unclear, however, whether people actually answer questions randomly. Several studies have found that responses to items about obscure or fictitious issues (on which respondents could not have had preexisting views) were not random but acted like meaningful opinions, i.e., were correlated with other attitudes and were stable over time (Bishop et al. 1983, Schuman & Presser 1981). Respondents who offered opinions appeared to do so only after constructing a meaning for the item

and then drawing on relevant predispositions in deciding how to answer (see also Strack et al. 1991, Tourangeau & Rasinski 1988).

This same process applies to questions about ordinary issues; even on familiar matters, individuals often cannot retrieve an answer to attitude questions and instead have to construct an answer from accessible predispositions (Sudman et al. 1996, Tourangeau et al. 2000). To the extent that respondents satisfice, this suggests that filters may reduce opinion giving by discouraging people from undertaking the cognitive effort needed to formulate an answer based on their preexisting attitudes.

How then does filtering reduce opinion giving—by eliciting fewer true attitudes or fewer nonattitudes? If filtering reduces nonattitudes, not true opinions, then indicators of data quality (e.g., temporal stability) should be higher with filtered items than with standard versions of the same questions. In three experiments, Krosnick et al. (2002) found no support for this hypothesis. There was essentially no difference in data quality between filtered and standard versions in any of their experiments. McClendon & Alwin (1993) also found no evidence that filtered questions improve reliability. If further research confirms this, then, as a general rule, it may be best to avoid filters and instead supplement direction-of-opinion measures with follow-up items on other attitudinal dimensions, such as salience and intensity.

## Questions That Involve Agreement: Acquiescence

Posing questions as statements to be agreed or disagreed with is among the most common formats found in attitude surveys, yet at the same time, it is the most controversial method of asking questions. On the one hand, agree-disagree items are simple to construct and easy to answer. On the other hand, they encourage acquiescence, the tendency to agree irrespective of item content.

Much of the literature on the correlates of acquiescence focuses on personality traits, although the key finding for general population surveys concerns the role of cognitive skills (for a comprehensive review, see Krosnick & Fabrigar 2003). Acquiescence occurs disproportionately among less-educated respondents—in recent American studies, among individuals who had not graduated from high school. As a result, the assumption of form-resistant correlations (the belief that wording changes may alter univariate distributions but not bivariate or multivariate distributions), which holds for many wording effects, does not extend to agree-disagree questions. Jackman (1973), for example, found that the use of agree-disagree questions affected the relationship between education and anti-Semitism. The negative correlation she observed with agree-disagree items was due to acquiescence; it disappeared with forced-choice questions.

To offset the effects of acquiescence, texts commonly recommend balancing the direction of agree-disagree items (by posing the same number of statements on each side of an issue). Yet this is unlikely to solve the problem, as it assumes the tendency to acquiesce is constant across items (and even if that were true, it is not clear why individuals who acquiesce should be assigned scores at the middle of the scale). Consequently, some researchers have counseled against the use of

these items and in favor of forced-choice questions (Converse & Presser 1986). For example, "Do you agree or disagree that most men are better suited emotionally for politics than are most women?" could be replaced by "Would you say that most men are better suited emotionally for politics than are most women, that men and women are equally suited, or that women are better suited than men in this area?" Similarly, the agree-disagree statement "Management lets employees know how their work contributes to the agency's goals" could be replaced by "Some people think management lets employees know how their work contributes to the agency's goals. Other people think management does not let employees know how their work contributes to the agency's goals. Which comes closer to how you feel?"

Forced-choice items may have advantages not only over agree-disagree items but over true-false items and yes-no items as well. Some evidence suggests that the reliability and validity of forced-choice questions is generally higher than that for either true-false or yes-no items, possibly because the latter approaches invite acquiescence by stating only one side of an issue (Krosnick & Fabrigar 2003).

Foregoing agree-disagree items may be problematic when investigators aim to replicate, extend, or otherwise make comparisons to previous studies that used such questions. In these instances, a useful approach involves a split-sample, administering the original agree-disagree item to a random set of cases, and a forced-choice version to the remainder. This allows for comparisons holding wording constant and also provides a gauge of acquiescence's impact on the results.

## TESTING AND EVALUATING QUESTIONS

It is an article of faith among survey researchers that pretesting should be an integral part of questionnaire development. For many decades, however, pretests involved only unstructured feedback from a few interviewers about a handful of interviews they conducted, sometimes supplemented by a review of the respondents' answers. The interviewers received no special training, and the respondents, who were not informed of the purpose of the pretest, were interviewed exactly as in a regular survey interview. Almost no research assessed the usefulness of this conventional practice, which appears better designed to reveal problems interviewers have with the questionnaire than problems respondents experience. As a result, to claim that a question had been pretested conveyed little information about its merits.

In the past 15 years, this has begun to change. Approaches specifically designed to identify questionnaire problems have been designed, and systematic inquiries into the nature of both the conventional and newer approaches have been undertaken (Cannell et al. 1989, Oksenberg et al. 1991, Presser & Blair 1994).

The two approaches that have received the most attention are cognitive interviews (Willis et al. 1991) and behavior coding (Fowler & Cannell 1996). Cognitive interviews are based on the work of Ericsson & Simon (1980), who asked experts to "think aloud" as they solved a complex problem, for instance, deciding what move to make in chess. This proved valuable in identifying the strategies experts used to perform a task, and the hope was that it would likewise identify the processes

respondents used to answer survey questions. But answering survey questions is unlike playing chess (or many other complex tasks) in that survey respondents frequently have access only to the results of their cognitive processes and not to the actual processes. People are therefore often unable to think aloud in cognitive interview pretests, and even when they are successful, thinking aloud may alter the way the task of answering the survey questions is performed. For these reasons, cognitive interviews for questionnaire testing have come to rely mainly on probes asked after each question is answered. These probes (e.g., "What did you think 'health care provider' meant?" and "What period of time were you thinking about when you answered that?") are similar to those used decades ago by methodologists such as Belson (1981, reporting on work from the 1960s) and Schuman (1966). In addition to questions about how cognitive interviews are best conducted, much still remains to be learned about how the data from these interviews should be analyzed.

Whereas cognitive interviews depart substantially from the conventional pretest, behavior coding supplements it. Codes are assigned to behaviors that occur during a conventional interview, for instance, "interviewer departs from question wording" and "respondent asks for clarification." Questions are then rated on how frequently they stimulate problematic behaviors. Developed initially by Cannell and his associates (Marquis & Cannell 1969), behavior coding's strength is its objective nature (which makes it reliable), but it provides little information about the cause of the problems it identifies unless coders augment their summary codes with detailed notes. Some respondent behaviors, such as giving adequate or qualified answers, are associated with the reliability of answers (Hess et al. 1999, Mathiowetz 1998). Dykema et al. (1997) also reported that qualified answers and a summary measure of several respondent behaviors were associated with less-accurate answers for measures about doctor visits, but contrary to expectation, interruptions by the respondent during the initial reading of an item and substantive changes in the question made by interviewers during the initial reading were associated with more accurate answers. Thus, additional work is needed to improve our understanding of the use of behavior coding to evaluate questions.

Various other approaches to test and evaluate questions have been developed, including respondent debriefings (Hess & Singer 1995), vignettes (Martin et al. 1991), and both manual and computerized evaluation schemes (Forsyth et al. 1992, Graesser et al. 1999). Schwarz & Sudman (1996) and the papers commissioned for the International Conference on Questionnaire Development, Evaluation and Testing Methods held in 2002 (Presser et al. 2003) provide an overview of these developments.

## CONCLUSION

For many years, there was little basis for quarreling with the title of Stanley Payne's 1951 classic. Asking questions was an art. Now, however, a body of work has accumulated that lays a foundation for a science of asking questions. Researchers can make decisions about some aspects of question wording informed by the results

of theoretically motivated experimental comparisons. Although asking questions will always involve an element of art, future research is likely to provide guidance for decisions about many other features of wording. The resulting improvements in survey measurement should facilitate progress in all areas of social science that make use of questionnaires.

ACKNOWLEDGMENTS

We thank Jennifer Dykema, Nancy Mathiowetz, Howard Schuman, Norbert Schwarz, and Roger Tourangeau for helpful comments on an earlier draft of this review.



LITERATURE CITED

Abelson RP, Loftus EF, Greenwald AG. 1992. Attempts to improve the accuracy of self-reports of voting. See Tanur 1992, pp. 138–53

Alwin DF. 1992. Information transmission in the survey interview: number of response categories and the reliability of attitude measurement. In *Sociological Methodology*, ed. PV Marsden, 22:83–118. Washington, DC: Am. Sociol. Assoc.

Alwin DF. 1997. Feeling thermometers versus 7-point scales: Which are better? *Sociol. Methods Res.* 25:318–40

Alwin DF, Krosnick JA. 1991. The reliability of survey attitude measurement: The influence of question and respondent attributes. *Sociol. Methods Res.* 20:139–81

Ayidiya SA, McClendon MJ. 1990. Response effects in mail surveys. *Public Opin. Q.* 54:229–47

Belli RF, Schwarz N, Singer E, Talarico J. 2000. Decomposition can harm the accuracy of behavioral frequency reports. *Appl. Cogn. Psychol.* 14:295–308

Belli RF, Shay WL, Stafford FP. 2001. Event history calendars and question list surveys: a direct comparison of interviewing methods. *Public Opin. Q.* 65:45–74

Belli RF, Traugott MW, Young M, McGonagle KA. 1999. Reducing vote overreporting in surveys: social desirability, memory failure, and source monitoring. *Public Opin. Q.* 63:90–108

Belson WA. 1981. *The Design and Understanding of Survey Questions*. Aldershot, Engl.: Gower

Bishop GF, Hippler H-J, Schwarz N, Strack F. 1988. A comparison of response effects in self-administered and telephone surveys. In *Telephone Survey Methodology*, ed. RM Groves, P Biemer, LE Lyberg, JT Massey, WL Nicholls II, J Waksberg, pp. 321–40. New York: Wiley

Bishop GF, Oldendick RW, Tuchfarber AJ, Bennett SE. 1983. Pseudo-opinions on public affairs. *Public Opin. Q.* 44:198–209

Blair E, Burton S. 1987. Cognitive processes used by survey respondents to answer behavioral frequency questions. *J. Consum. Res.* 14:280–88

Blair E, Sudman S, Bradburn NM, Stocking C. 1977. How to ask questions about drinking and sex: response effects in measuring consumer behavior. *J. Mark. Res.* 14:316–21

Bohrnstedt GW. 1983. Measurement. In *Handbook of Survey Research*, ed. PH Rossi, JD Wright, AB Anderson, pp. 70–114. Orlando: Academic

Bradburn NM, Miles C. 1979. Vague quantifiers. *Public Opin. Q.* 43:92–101

Bradburn NM, Rips LJ, Shevell SK. 1987. Answering autobiographical questions: the

impact of memory and inference on surveys. *Science* 236:157–61

Brown NR, Rips LJ, Shevell SK. 1985. The subjective dates of natural events in very long term memory. *Cogn. Psychol.* 17:139–77

Burton S, Blair E. 1991. Task conditions, response formulation processes, and response accuracy for behavioral frequency questions in surveys. *Public Opin. Q.* 55:50–79

Cannell CF, Miller PV, Oksenberg L. 1981. Research on interviewing techniques. In *Sociological Methodology*, ed. S Leinhardt, 11:389–437. San Francisco: Jossey-Bass

Cannell CF, Oksenberg L, Kalton G, Bischoping K, Fowler FJ. 1989. *New Techniques for Pretesting Survey Questions*. *Final Report*. *Grant No. HS 05616*. Natl. Cent. Health Serv. Res. Health Care Technol. Assess.

Cliff N. 1959. Adverbs as multipliers. *Psychol. Rev.* 66:27–44

Collins M. 1980. Interviewer variability: a review of the problem. *J. Mark. Res. Soc.* 22:77–95

Conrad FG, Brown NR, Cashman ER. 1998. Strategies for estimating behavioural frequency in survey interviews. *Memory* 6:339–66

Conrad FG, Schober MF. 2000. Clarifying question meaning in a household telephone survey. *Public Opin. Q.* 64:1–28

Converse JM, Presser S. 1986. *Survey Questions: Handcrafting the Standardized Questionnaire*. Beverly Hills, CA: Sage

Converse P. 1964. *The Nature of Belief Systems in Mass Publics. In Ideology and Discontent*, ed. D Apter, pp. 206–61. New York: Free Press

Cox EP. 1980. The optimal number of response alternatives for a scale: a review. *J. Mark. Res.* 17:407–22

Czaja R, Blair J, Bickart B, Eastman E. 1994. Respondent strategies for recall of crime victimization incidents. *J. Off. Stat.* 10:257–76

Dillman DA, Mason RG. 1984. *The influence of survey method on question response*. Presented at Annu. Meet. Am. Assoc. Public Opin. Res., Delavan, WI

Dobson KS, Mothersill KJ. 1979. Equidistant categorical labels for construction of Likert type scales. *Percept. Mot. Skills* 49:575–80

Dominitz J, Manski CF. 1997. Perceptions of economic insecurity: evidence from the survey of economic expectations. *Public Opin. Q.* 61:261–87

Dykema J, Lepkowski JM, Blixt S. 1997. The effect of interviewer and respondent behavior on data quality: analysis of interaction coding in a validation study. See Lyberg et al. 1997, pp. 287–310

Dykema J, Schaeffer NC. 2000. Events, instruments, and reporting errors. *Am. Sociol. Rev.* 65:619–29

Ericsson KA, Simon HA. 1980. Verbal reports as data. *Psychol. Rev.* 87:215–51

Forsyth B, Lessler JT, Hubbard M. 1992. Cognitive evaluation of the questionnaire. See Turner et al. 1992, pp. 13–52

Fowler FJ Jr. 1992. How unclear terms affect survey data. *Public Opin. Q.* 56:218–31

Fowler FJ Jr, Cannell CF. 1996. Using behavioral coding to identify cognitive problems with survey questions. See Schwarz & Sudman 1996, pp. 15–36

Freedman D, Thornton A, Camburn D, Alwin DF, Young-DeMarco L. 1988. The life-history calendar: a technique for collecting retrospective data. In *Sociological Methodology*, ed. C Clogg, 18:37–68. Washington, DC: Am. Sociol. Assoc.

Fuchs M. 2002. The impact of technology on interaction in computer-assisted interviews. See Maynard et al. 2002, pp. 471–91

Graesser AC, Kennedy T, Wiemer-Hastings P, Ottati P. 1999. The use of computational cognitive models to improve questions on surveys and questionnaires. See Sirken et al. 1999, pp. 199–216

Groves RM, Kahn RL. 1979. *Surveys by Telephone: A National Comparison With Personal Interviews*. New York: Academic

Hess J, Singer E. 1995. *The role of respondent debriefing questions in questionnaire development*. Presented at Annu. Meet. Am. Assoc. Public Opin. Res., Fort Lauderdale, FL

Hess J, Singer E, Bushery JM. 1999. Predicting

test-retest reliability from behavior coding. *Int. J. Public Opin. Res.* 11:346–60

Huttenlocher J, Hedges LV, Bradburn NM. 1990. Reports of elapsed time: bounding and rounding processes in estimation. *J. Exp. Psychol. Learn. Mem. Cognit.* 16:196–213

Igou E, Bless H, Schwarz N. 2002. Making sense of standardized survey questions: the influence of reference periods and their repetition. *Commun. Monogr.* 69:179–87

Jackman MR. 1973. Education and prejudice or education and response set? *Am. Sociol. Rev.* 38:327–39

Jobe JB, Pratt WF, Tourangeau R, Baldwin AK, Rasinski KA. 1997. Effects of interview mode on sensitive questions in a fertility survey. See Lyberg et al. 1997, pp. 311–52

Klockars AJ, Yamagishi M. 1988. The influence of labels and positions in rating scales. *J. Educ. Meas.* 25:85–96

Klopfer FJ, Madden TM. 1980. The middlemost choice on attitude items: ambivalence, neutrality or uncertainty? *Personal. Soc. Psychol. Bull.* 6:91–101

Knauper B. 1998. Filter questions and question interpretation: presuppositions at work. *Public Opin. Q.* 62:70–78

Krosnick JA. 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. Special issue: cognition and survey measurement. *Appl. Cogn. Psychol.* 5:213–36

Krosnick JA, Berent MK. 1993. Comparisons of party identification and policy preferences: the impact of survey question format. *Am. J. Polit. Sci.* 37:941–64

Krosnick JA, Fabrigar LR. 2003. *Designing Questionnaires to Measure Attitudes*. New York: Oxford Univ. Press. In press

Krosnick JA, Holbrook AL, Berent MK, Carson RT, Hanemann WM, et al. 2002. The impact of 'no opinion' response options on data quality: non-attitude reduction or invitation to satisfice? *Public Opin. Q.* 66:371–403

Lee L, Brittingham A, Tourangeau R, Rasinski K, Willis G, et al. 1999. Are reporting errors due to encoding limitations or retrieval failure? *J. Appl. Cogn. Psychol.* 13:43–63

Loftus EF, Smith KD, Klinger MR, Fiedler J. 1992. Memory and mismemory for health events. See Tanur 1992, pp. 102–37

Lyberg L, Biemer P, Collins M, de Leeuw E, Dippo C, et al. 1997. *Survey Measurement and Process Quality.* New York: Wiley

Mangione TW Jr, Fowler FJ, Louis TA. 1992. Question characteristics and interviewer effects. *J. Off. Stat.* 8:293–307

Marquis KH, Cannell CF. 1969. *A study of interviewer-respondent interaction in the urban employment survey.* Survey Res. Cent., Inst. Soc. Res., Univ. Mich.

Martin E, Campanelli P, Fay RE. 1991. An application of rasch analysis to questionnaire design: using vignettes to study the meaning of 'work' in the current population survey. *The Stat.* 40:265–76

Mathiowetz NA. 1998. Respondent expressions of uncertainty: data source for imputation. *Public Opin. Q.* 62:47–56

Mathiowetz NA, Duncan GJ. 1988. Out of work, out of mind: response errors in retrospective reports of unemployment. *J. Bus. Econ. Stat.* 6:221–29

Maynard DW, Houtkoop-Steenstra H, Schaeffer NC, van der Zouwen J, eds. 2002. *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*. New York: Wiley

McClendon MJ, Alwin DF. 1993. No opinion filters and attitude measurement reliability. *Sociol. Methods Res.* 21:438–64

Means B, Loftus EF. 1991. When personal history repeats itself: decomposing memories for recurring events. *Appl. Cogn. Psychol.* 5:297–318

Means B, Nigam A, Zarrow M, Loftus EF, Donaldson MS. 1989. *Autobiographical memory for health-related events*. Cogn. Surv. Meas., Ser. 6, No. 2. Rockville, MD: Natl. Cent. Health Stat.

Menon G. 1993. The effects of accessibility of information in memory on judgments of behavioral frequencies. *J. Consum. Res.* 20:431–40

Menon G. 1997. Are the parts better than the whole? The effects of decompositional

questions on judgments with frequent behaviors. *J. Mark. Res.* 34:335–46

Miller GA. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* 63:81–97

Moore JC, Moyer L. 1998. Questionnaire design effects on interview outcomes. In *American Statistical Association Proceedings of the Section on Survey Research Methods*, pp. 851–56. Washington, DC: Am. Stat. Assoc.

Neter J, Waksberg J. 1964. A study of response errors in expenditures data from household interviews. *J. Am. Stat. Assoc.* 59:18–55

Oksenberg L, Beebe TJ, Blixt S, Cannell C. 1992. *Research on the Design and Conduct of the National Medical Expenditure Survey Interviews, Final Report*. Ann Arbor, MI: Surv. Res. Cent., Inst. Soc. Res.

Oksenberg L, Cannell CF, Kalton G. 1991. New strategies for pretesting survey questions. *J. Off. Stat.* 7:349–65

O'Muircheartaigh CA, Gaskell G, Wright DB. 1995. Weighing anchors: verbal and numeric labels for response scales. *J. Off. Stat.* 11:295–307

O'Muircheartaigh CA, Krosnick JA, Helic A. 1999. *Middle alternatives, acquiescence, and the quality of questionnaire data*. Presented at Annu. Meet. Am. Assoc. Public Opin. Res., Fort Lauderdale, FL

Payne SL. 1951. *The Art of Asking Questions*. Princeton, NJ: Princeton Univ. Press

Pohl NF. 1981. Scale considerations in using vague quantifiers. *J. Exp. Educ.* 49:235–40

Presser S. 1990. Can changes in context reduce vote overreporting in surveys? *Public Opin. Q.* 54:586–93

Presser S, Blair J. 1994. Survey pretesting: Do different methods produce different results? In *Sociological Methodology*, ed. PV Marsden, pp. 73–104. New York: Blackwell

Presser S, Rothgeb J, Couper M, Lessler J, Martin E, et al. eds. 2003. *Methods for Testing and Evaluating Survey Questionnaires*. New York: Wiley. In press

Presser S, Stinson L. 1998. Data collection mode and social desirability bias in self-reported religious attendance. *Am. Sociol. Rev.* 63:137–45

Presser S, Traugott M. 1992. Little white lies and social science models: correlated response errors in a panel study of voting. *Public Opin. Q.* 56:77–86

Ross M, Conway M. 1986. Remembering one's own past: the construction of personal histories. In *Handbook of Motivation and Cognition: Foundations of Social Behavior*, ed. R Sorrentino, ET Higgins, pp. 122–44. Chichester, Engl.: Wiley

Schaeffer NC. 1991. Hardly ever or constantly? Group comparisons using vague quantifiers. *Public Opin. Q.* 55(3):395–423

Schaeffer NC. 1994. Errors of experience: response errors in reports about child support and their implications for questionnaire design. In *Autobiographical Memory and the Validity of Retrospective Reports*, ed. N Schwarz, S Sudman, pp. 141–60. New York: Springer-Verlag

Schaeffer NC. 2000. Asking questions about threatening topics: a selective overview. In *The Science of Self-Report: Implications for Research and Practice*, ed. AA Stone, JS Turkkan, CA Bachrach, JB Jobe, HS Kurtzman, VS Cain, pp. 105–22. Mahwah, NJ: Erlbaum

Schaeffer NC, Barker K. 1995. *Alternative methods of presenting bi-polar scales in telephone interviews: 1 to 7 vs -3 to +3 and neutral vs ambivalent*. Presented at Annu. Meet. Am. Assoc. Public Opin. Res., Fort Lauderdale, FL

Schaeffer NC, Bradburn NM. 1989. Respondent behavior in magnitude estimation. *J. Am. Stat. Assoc.* 84:402–13

Schaeffer NC, Charng H-W. 1991. Two experiments in simplifying response categories: intensity ratings and behavioral frequencies. *Sociol. Perspect.* 34:165–82

Schaeffer N, Dykema J. 2003. A multiple method approach to improving the clarity of closely related concepts. See Presser et al. 2003

Schaeffer NC, Guzman L. 1999. *Interpreting reference periods*. Presented at Annu.

Meet. Am. Assoc. Public Opin. Res., St. Pete Beach, FL

Schaeffer NC, Maynard DW. 1996. From paradigm to prototype and back again: interactive aspects of cognitive processing in standardized survey interviews. See Schwarz & Sudman 1996, pp. 65–88

Schaeffer NC, Maynard DW. 2002. Occasions for intervention: interactional resources for comprehension in standardized survey interviews. See Maynard et al. 2002, pp. 261–80

Schober MF, Conrad FG. 1997. Does conversational interviewing reduce survey measurement error? *Public Opin. Q.* 61:576–602

Schuman H. 1966. The random probe: a technique for evaluating the validity of closed questions. *Am. Sociol. Rev.* 31:218–22

Schuman H, Ludwig J. 1983. The norm of even-handedness in surveys as in life. *Am. Sociol. Rev.* 48:112–20

Schuman H, Presser S, eds. 1981. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context.* New York: Academic

Schuman H, Scott J. 1987. Problems in the use of survey questions to measure public opinion. *Science* 236:957–59

Schwarz N. 1994. Judgment in social context: biases, shortcomings, and the logic of conversation. *Adv. Exp. Soc. Psychol.* 26:123–62

Schwarz N, Grayson C, Knauper B. 1998. Formal features of rating scales and the interpretation of question meaning. *Int. J. Public Opin. Res.* 10:177–83

Schwarz N, Hippler HJ, Deutsch B, Strack F. 1985. Response scales: effects of category range on reported behavior and comparative judgments. *Public Opin. Q.* 49:388–95

Schwarz N, Knauper B, Hippler HJ, Noelle-Neumann E, Clark L. 1991. Rating scales: Nueric values may change the meaning of scale labels. *Public Opin. Q.* 55:570–82

Schwarz N, Strack F, Mai HP. 1991. Assimilation and contrast effects in part-whole question sequences: a conversational logic analysis. *Public Opin. Q.* 55:3–23

Schwarz N, Strack F, Muller G, Chassein B. 1988. The range of response alternatives may determine the meaning of the question: further evidence on informative functions of response alternatives. *Soc. Cogn.* 6:107–17

Schwarz N, Sudman S, eds. 1996. *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research.* San Francisco: Jossey-Bass

Sirken MG, Herrmann DJ, Schechter S, Schwarz N, Tanur JM, Tourangeau R, eds. 1999. *Cognition and Survey Research.* New York: Wiley

Strack F, Schwarz N, Wanke M. 1991. Semantic and pragmatic aspects of context effects in social and psychological research. *Soc. Cogn.* 9:111–25

Suchman L, Jordan B. 1990. Interactional troubles in face-to-face survey interviews. *J. Am. Stat. Assoc.* 85:232–53

Sudman S, Bradburn NM, Schwarz NM. 1996. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology.* San Francisco: Jossey-Bass

Sudman S, Finn A, Lannom L. 1984. The use of bounded recall procedures in single interviews. *Public Opin. Q.* 48:520–24

Tanur JM, ed. 1992. *Questions About Questions: Inquiries into the Cognitive Bases of Surveys.* New York: Russell Sage Found.

Tourangeau R. 1984. Cognitive sciences and survey methods. In *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, ed. T Jabine, M Straf, J Tanur, R Tourangeau, pp. 73–100. Washington, DC: Natl. Acad. Press

Tourangeau R, Rasinski K. 1988. Cognitive processes underlying context effects in attitude measurement. *Psychol. Bull.* 103:299–314

Tourangeau R, Rasinski K, Bradburn N. 1991. Measuring happiness in surveys: a test of the subtraction hypothesis. *Public Opin. Q.* 55:255–66

Tourangeau R, Rips LJ, Rasinski K. 2000. The psychology of survey response. Cambridge, Engl.: Cambridge Univ. Press

Tourangeau R, Smith TW. 1996. Asking sensitive questions: the impact of data collection mode, question format, and question context. *Public Opin. Q.* 60:275–304

Turner CF, Forsyth BH, O'Reilly JM, Cooley PC, Smith TK, et al. 1998. Automated self-interviewing and the survey measurement of sensitive behaviors. In *Computer Assisted Survey Information Collection*, ed. MP Couper, RP Baker, J Bethlehem, CZF Clark, J Martin, et al., pp. 455–74. New York: Wiley

Turner CF, Lessler JT, Gfroerer JC, eds. 1992. *Survey Measurement of Drug Use*. Rockville, MD: Nat. Inst. Drug Abus., US Dep. Health Hum. Serv.

Turner CF, Martin E, eds. 1984. *Surveying Subjective Phenomena*. New York: Russell Sage Found.

Tversky A, Kahneman D. 1973. Availability: a heuristic for judging frequency and probability. *Cogn. Psychol.* 5:207–32

Willis GB, Royston P, Bercini D. 1991. The use of verbal report methods in the development and testing of survey questionnaires. *Appl. Cogn. Psychol.* 5:251–67

Winkielman P, Knauper B, Schwarz N. 1998. Looking back at anger: Reference periods change the interpretation of emotion frequency questions . *J. Personal. Soc. Psychol.* 75:719–28